Research Article | Vol. 13, No. 4 / 1 Apr 2022 / Biomedical Optics Express 1968
Biomedical Optics EXPRESS

Automatic zoning for retinopathy of prematurity with semi-supervised feature calibration adversarial learning

YUANYUAN PENG,[1] ZHONGYUE CHEN,[1] WEIFANG ZHU,[1] FEI SHI,[1] MENG WANG,[1] YI ZHOU,[1] DAOMAN XIANG,[3] XINJIAN CHEN,[1,2,4] AND FENG CHEN[3,5]

[1]*MIPAV Lab, School of Electronics and Information Engineering, Soochow University, Suzhou, Jiangsu 215006, China*
[2]*State Key Laboratory of Radiation Medicine and Protection, Soochow University, Suzhou 215123, China*
[3]*Guangzhou Women and Children's Medical Center, Guangzhou 510623, China*
[4]*xjchen@suda.edu.cn*
[5]*eyeguangzhou@126.com*

Abstract: Retinopathy of prematurity (ROP) is an eye disease, which affects prematurely born infants with low birth weight and is one of the main causes of children's blindness globally. In recent years, there are many studies on automatic ROP diagnosis, mainly focusing on ROP screening such as "Yes/No ROP" or "Mild/Severe ROP" and presence/absence detection of "plus disease". Due to the lack of corresponding high-quality annotations, there are few studies on ROP zoning, which is one of the important indicators to evaluate the severity of ROP. Moreover, how to effectively utilize the unlabeled data to train model is also worth studying. Therefore, we propose a novel semi-supervised feature calibration adversarial learning network (SSFC-ALN) for 3-level ROP zoning, which consists of two subnetworks: a generative network and a compound network. The generative network is a U-shape network for producing the reconstructed images and its output is taken as one of the inputs of the compound network. The compound network is obtained by extending a common classification network with a discriminator, introducing adversarial mechanism into the whole training process. Because the definition of ROP tells us where and what to focus on in the fundus images, which is similar to the attention mechanism. Therefore, to further improve classification performance, a new attention mechanism based feature calibration module (FCM) is designed and embedded in the compound network. The proposed method was evaluated on 1013 fundus images of 108 patients with 3-fold cross validation strategy. Compared with other state-of-the-art classification methods, the proposed method achieves high classification performance.

1. Introduction

Retinopathy of prematurity (ROP) is retinal vascular proliferative blindness disease, which frequently occurs in premature babies with low birth weight (less than 1500 g) and accounts for about 19% of the causes of blindness in children worldwide [1,2]. It is reported that in 2010, about 184700 preterm infants in the world had ROP with different degrees, and about 20000 of them had severe visual impairment or blindness [1]. In addition, ROP will lead to strabismus in 30% of the premature infants without timely treatment [3].

Early diagnosis and timely treatment of ROP can effectively prevent the growth of the abnormal vessels, thus prevent the disease turning into blindness [4,5]. The diagnosis and analysis of ROP is usually based on the premature infants' retinal fundus images obtained by using RetCam3, which is a digital retinal camera with high-quality and wide-angle image [6]. According to the International Classification of ROP (ICROP) [7–9], the diagnosis of ROP is subclassified by

#447224 https://doi.org/10.1364/BOE.447224
Journal © 2022 Received 1 Nov 2021; revised 5 Jan 2022; accepted 9 Feb 2022; published 9 Mar 2022

Biomedical Optics EXPRESS

zone, stage and plus disease. In this study, we focus on studying the automatic recognition of three zones of ROP, which is defined according to the location of the symptom of ROP relative to the optic disc and macular, as shown in Fig. 1. The detailed definition of Zone I to III are listed in Table 1 and Fig. 2 shows the examples of ROP with different zones.

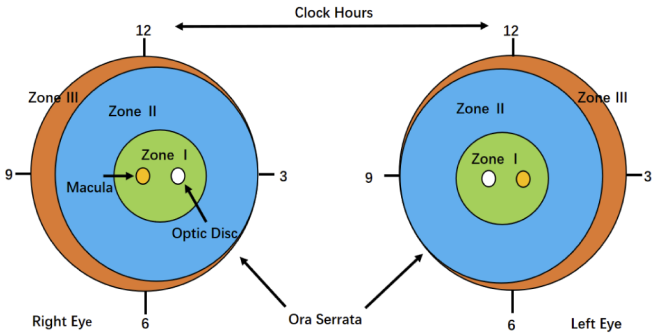

Fig. 1. Standard form for documenting zone.

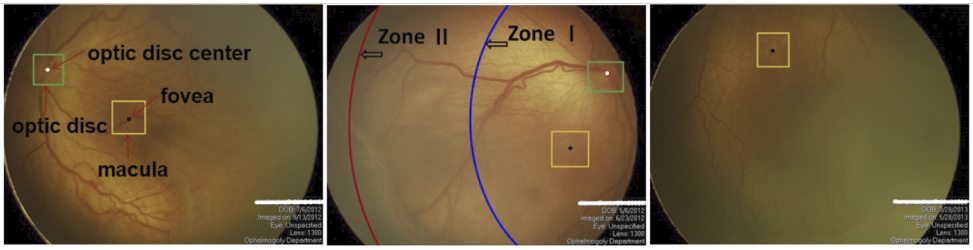

Fig. 2. Examples of zone I, zone II and zone III. The optic disc and macula are in the green and yellow boxes respectively, while the white and black dots indicate the center of the optic disc and the fovea of the macula respectively. (a) Zone I. (b) Zone II. (c) Zone III.

Table 1. Definition of zone I to III of ROP.

Zone	Definition
I	A circular area with a radius of twice the distance from the center of the optic disc to the fovea of the macula.
II	An annular area with a radius of the distance from the optic disc to the nasal serrated margin except zone I.
III	The remaining crescent shaped areas outside zone I and II vitreous.

In the past, many related studies on automated or semi-automated methods for ROP diagnosis are mainly for the recognition of plus disease, which is characterized by dilation and tortuosity of retinal vessels. For example, Wittenberg et al. used "ROPTool" system to assist the ophthalmologists in diagnosing plus disease by calculating tortuosity of vessels, which was a manual or semi-automated process [10]. Ataer-Cansizoglu et al. used support vector machine to learn the best relationship between features and diagnosis for automatic diagnosis of three types of plus diseases that performed as well as experts [11]. However, the clinical application of the system is limited due to the need of manually tracked and segmented vessels as input, which is a time-consuming and laborious work. Worrall et al. used pre-trained GoogLeNet with approximate Bayesian posterior for fully automated plus disease diagnosis, which was the first time to use deep convolutional network for ROP diagnosis [12]. Brown et al. proposed an "i-ROP" deep learning

Biomedical Optics EXPRESS

system for 3-level classification of plus disease, which achieved a high classification accuracy [13]. Meanwhile, there are also some studies for ROP screening, ROP severity grading, ROP staging and ROP zoning [14–19]. For example, a deep learning system called "DeepROP" was developed for the detection of ROP, which was based on Inception-V2 pre-trained on ImageNet and achieves high classification accuracy [14,15]. Zhang et al used VGG16 pre-trained on ImageNet for automatic ROP screening [16]. VGG16 and ResNet50 pre-trained on ImageNet were used for the automated recognition of aggressive posterior retinopathy of prematurity (AP-ROP), which is characterized by severe vasodilation and distortion of the posterior pole of the retina [17]. A joint segmentation and multi-instance learning based CNNs was proposed for the automatic stage of ROP, which only involved 4-level ROP staging [18]. Our previous work included automated ROP screening by using ResNet18 pre-trained on ImageNet with attention mechanism [19], automatic ROP zoning by using ImageNet pre-trained DenseNet121with attention mechanism and deep supervision strategy [20] and automatic ROP staging with transfer learning, feature fusion and ordinal classification strategy [21]. In addition, Zhao et al. developed a deep learning framework to automatically identify zone I, which can draw the boundary of zone I on the fundus images as a diagnostic aid [22]. There are some limitations in [22]. First, the recognition of zone I is based on the detection of optic disc and macula, so the algorithm cannot be used to analyze the retinal fundus images without optic disc and macula. Second, the performance of this method depends on the detection accuracy of optic disc and macula. However, due to the incomplete development of the macula in the newborn, the macular structure is not obvious in the corresponding fundus images, which may lead to the low recognition accuracy of the macula, and then affect the recognition accuracy of ROP zone I. Finally, the algorithm only realizes the automatic recognition of zone I and does not involve the automatic recognition of zone II and zone III, which is also important for the assessment of ROP severity. Recently, Ranjana et al. proposed a method using U-Net and circle Hough transform to detect zones I, II and III, which involves optic disc and blood vessel segmentation [23]. In their method, macula's location was determined according to the Refs. [24] and [25] and repeated verification by senior ROP specialists. In addition, the detection of zone III in [23] is limited.

In conclusion, deep learning holds promise for automated and objective diagnosis of ROP in digital fundus images. However, there are still some challenges in achieving accurate ROP zoning. On the one hand, compared to ROP screening and ROP staging task, which generally have relatively more labeled data, especially ROP screening, the data for the ROP zoning task is limited and the corresponding annotation is difficult to be obtained. On the other hand, different from ROP screening and ROP staging task, which only need to correctly identify lesion and the difference between lesions, ROP zoning not only need to pay attention to lesion but also the positional relationship between lesion and optic disc and macula. As we all know, a deep neural network usually needs large numbers of images with corresponding high-quality annotations, which is time-consuming and requires large amounts of expert knowledge. Therefore, traditional CNN-based classification networks such as VGGNet, GoogleNet, DenseNet and ResNet may be ineffective. Aiming at the first challenge, semi-supervised deep learning algorithms have attracted our attention, which can combine labeled data sets with unlabeled data sets. Recently, many related works have used semi-supervised methods based on generative adversary networks (GANs) and achieved good classification performance in different medical image classification tasks [26–29]. In these networks, a discriminator of GAN and a classifier are unified into a single network, and the common generator takes noise as input to fit the real data statistical and produces the fake image as real as possible. Considering the simplicity of the training and inspired by the image reconstruction and transformation with GAN [29–32], the generative network in our method is a U-shape network, which takes original unlabeled fundus images of ROP as input instead of random noise and reconstructs input images as much as possible. Meanwhile, the discriminator strives to distinguish between input images and reconstructed images. In addition,

Biomedical Optics EXPRESS

the definition of ROP shown in Table 1 tells us where and what to focus on in the fundus images, which is similar to attention mechanism [33]. Meanwhile, previous studies have also shown that attention mechanisms can help to learn intermediate features to improve the performance of convolutional neural networks [34,35]. Therefore, focusing on the second challenge and inspired by Dual Attention Network (DANet) and Squeeze-and-Excitation block (SE block) [34,35], we propose a new attention module named feature calibration module (FCM), which can adaptively calibrate the features of spatial and channel dimensions and promote feature learning. To sum up, we apply semi-supervised method based on adversarial learning and attention mechanism for automatic ROP zoning with 3-level in this study, which can achieve good performance with labeled data and unlabeled data. The main contributions of this paper can be summarized as follows:

(1) A novel semi-supervised classification network based on adversarial learning is proposed for 3-level ROP zoning, introducing unlabeled data to assist classifier training and improving the generalization ability of the model. It is the first time to employ the semi-supervised learning method for ROP zoning.

(2) A novel feature calibration module is proposed, which can adaptively calibrate both the features of spatial and channel dimensions to promote feature learning, and further improve the accuracy of ROP zoning.

(3) Extensive experiments are conducted to evaluate the effectiveness of the proposed method. Experimental results show that the proposed method outperforms other state-of-the-art classification methods in ROP zoning task.

The remainder of this paper is organized as follows: The proposed method for automatic ROP zoning is introduced in Section 2. Section 3 presents the experimental results in detail. In section 4, we conclude this paper and suggest future work.

2. Methodology

2.1. Overview

Our proposed SSFC-ALN based ROP zoning framework is shown in Fig. 3(a), which consists of a generative network and a compound network. The generative network is a U-shape encode and decode network, which is used to reconstruct the original input fundus image. Supposing the total fundus image data set is $D_{T=}\{D_L, D_U\}$, $D_L = \{(x_i, y_i)\}$ and $D_u = \{x_i\}$ represent the labeled data and the unlabeled data respectively, in which x_i is the original fundus image and y_i is the corresponding label of x_i. In addition, D_R is the reconstructed fundus image. The proposed SSFC-ALN based ROP zoning framework is optimized as follows:

(1) Generative network initialization: the generative network is randomly initialized through normal distribution.

(2) Generative network: we fix the compound network first and then train the generator once by minimizing the L_G^{unsup}, which will be described in detail below.

(3) Compound network initialization: the weights of DenseNet121 pre-trained on the ImageNet are used to initialize the compound network.

(4) Compound network training: We fix the generative network and train the compound network twice by minimizing $L_{COM}^{semi-sup}$, which will be described in detail below.

(5) Iterate the training process of (2) and (4) for several epochs until the compound network converges.

Biomedical Optics EXPRESS

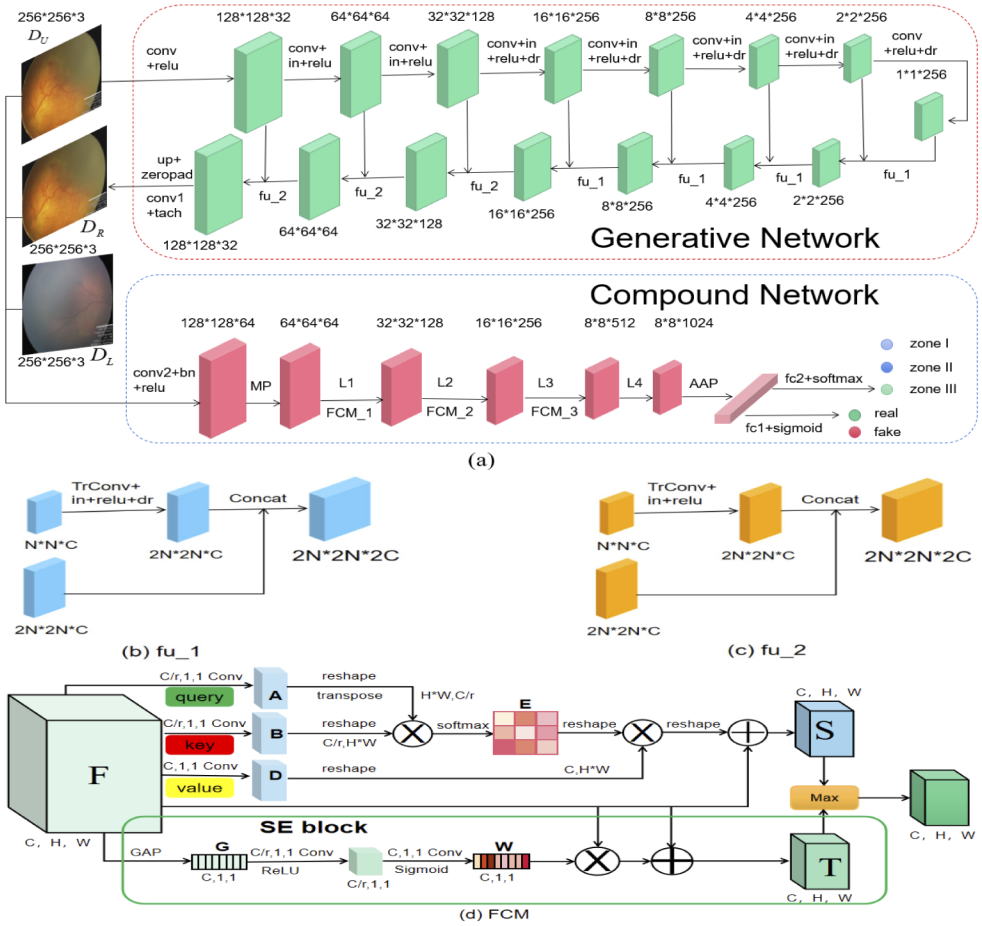

Fig. 3. Overview of the proposed SSFC-ALN based ROP zoning framework. The generative network for image reconstruction is in the red dotted box, where 'fu_1' and 'fu_2' are skip fusion operators as shown in (b) and (c) respectively, where 'TrConv', 'in', 'dr' and 'Concat' represent transpose convolution operator, instance normalization operator, dropout operator and concatenation operator. The compound network consisting of a classifier and a discriminator is in the blue dotted box, where 'MP' and 'AAP' represent max pooling operator and adaptive average pooling operator, 'L' represents multiple stacked dense connection modules and 'FCM' represents the proposed feature calibration module as shown in (d).

2.2. SSFC-ALN framework

Similar to the previous GAN-based methods [26–28], a classifier and a discriminator of GAN are unified into a single network and both of them share the same convolutional feature extractor. For convenience, we call this network as compound network. Original GAN [36] generates images based on random noise, but the randomness of its input also causes the uncontrollability of the output. To handle this problem, cGAN improves the controllability of the generated data by introducing some specific input signals as control conditions, such as class labels or data from other modalities [37]. Meanwhile, many previous medical image processing works also directly adopted images as input to generate target domain images [38–40], which attract our attention and prompt us to use the original image as the input of the generative network for image

Biomedical Optics EXPRESS

reconstruction. As can be seen from Fig. 3(a), our framework consists of two subnetworks, one for producing fake images, and the other for classification. The former is a generative network for images reconstruction with U-shape architecture, which is an end-to-end (image in, image out) network consisting of a series of convolutional layers, transpose convolutional layers, instance normalized layers and dropout layers, and contains a contraction path for capturing semantics and a symmetric extension path for precise positioning. The latter is a compound network using as classifier and discriminator, respectively. In addition, considering the convergence speed and memory overhead, we exploit DenseNet121 [41] pre-trained on ImageNet [42] as the feature extractor for the compound network followed by two different fully connected layers, which are modified according to the outputs two-dimension values. As can be seen from Fig. 3(a), compared with the general classification networks, we introduce semi-supervised learning through the generative network in the training process, which can introduce the adversarial mechanism to enhance the feature extraction ability of the classifier. Especially, total parameters of the proposed SSFC-ALN are about 20.9923M in the training stage, while in the test stage, only the trained compound network with about 7.3885M parameters is used to realize ROP zoning. In addition, our framework schematic depicted in Fig. 3(a) is theoretically easy to deploy to other common convolutional neural networks, such as ResNet34, ResNet50, VGG16, Inceptions et al [43–45].

2.3. Feature calibration module (FCM)

The definition of zones is according to the optic disc and macula location with the clinical lesions features such as ridge and blood vessels, whose zoning characteristics are depicted in Table 1. As can be seen from Table 1, the definition of ROP tells us where and what to focus on in the fundus images, which is similar to attention mechanism [33]. Therefore, the attention mechanism is adopted to enhance feature learning and improve the performance of ROP zoning.

The essence of attention mechanism is to locate the information of interest and suppress useless information [34,35], which can be mainly divided into three types: spatial attention module, channel attention module and spatial and channel mixed attention module. Many previous studies have shown that applying attention mechanism to convolutional neural network can increase its representation power to focus on important features, suppress the irrelevant ones and improve the performance in computer vision tasks [34,35], [46–51]. For example, Hu et al. [35] proposed a compact module named Squeeze-and-Excitation module to enhance the representational power of the network by modeling channel-wise relationship [35]. Fu et al. proposed a Dual Attention Network (DANet) to capture rich context dependence based on self-attention mechanism, which emphasizes meaningful features along channel and spatial axes [34].

To better learn the potential relationship between the location and lesion of ROP zoning, inspired by many previous successful applications of attention mechanisms, a novel attention module named feature calibration module (FCM) is proposed and embedded into the compound network, which allows the network to recalibrate features from the two dimensions of space and channel, and fuses the two calibrated features to obtain more expressive and effective feature. Specifically, given a feature map $\mathrm{F} \in \mathbb{R}^{C,H,W}$ as input, FCM can generate a spatial attention map $\mathrm{S} \in \mathbb{R}^{C,H,W}$ and a channel attention map $\mathrm{T} \in \mathbb{R}^{C,H,W}$ in parallel as shown in Fig. 3(d). Then, we fuse the two features to obtain better representations as illustrated in Eq. (1).

$$N = S \odot T \tag{1}$$

where $\odot$ denotes a feature fusion operator. In this paper, we adopt max fusion operator to obtain the maximum value of the features S and T.

Spatial attention module. Spatial attention can be understood as where the neural network pays attention to. Through spatial attention mechanism, the spatial information in the original feature can be transformed into another space and the key information can be retained**.** As shown in Fig. 3(d), given an input feature $\mathrm{F} \in \mathbb{R}^{C,H,W}$, we first feed it into three convolution layers

Biomedical Optics EXPRESS

with the size of 1*1 to generate three new feature maps A, B and D respectively, which are similar to the three branches of self-attention (query, key and value) [34,52] ($\{A, B\} \in \mathbb{R}^{C/r,H,W}$ and $D \in \mathbb{R}^{C,H,W}$). Second, we reshape A and B to $\mathbb{R}^{C/r,H*W}$ and D to $\mathbb{R}^{C,H*W}$, where C, H and W represent the channel numbers, height and width of the input feature and r is compression ratio. Then, we do a matrix multiplication between the transpose of A and B, and use a softmax activation function to calculate the spatial attention map $E \in \mathbb{R}^{H*W,H*W}$ as illustrated in Eq. (2). After that, we do a matrix multiplication between D and the transpose of E and reshape the obtained result to $\mathbb{R}^{C,H,W}$. Finally, we do an element-wise summation between the original input feature F and the above result to obtain the final spatial attention output $S \in \mathbb{R}^{C,H,W}$ as illustrated in Eq. (3).

$$e_{ji} = \frac{exp(A_i \cdot B_j)}{\sum_{i=1}^{H*W} exp(A_i \cdot B_j)} \tag{2}$$

$$S_j = \sum_{i=1}^{H*W} (e_{ji} D_i) + F_j \tag{3}$$

where e_{ji} measures the influence of the i-th position on j-th position. The more similar feature representations of the two positions, the greater correlation between them.

Channel attention module. Channel attention can be understood as what the neural network focus on. Different from spatial attention mechanism, channel attention can adaptively recalibrate the characteristic response of channels by explicitly modeling the interdependence between channels [34]. The typical representative of channel attention modules is SE block, which uses two fully connected layers to learn the relationship between different channels, thereby introducing a large number of parameters and increasing the risk of overfitting. Therefore, to reduce the calculations, two 1×1 convolution kernels are used to replace the fully connected layers in this paper. As illustrated in Fig. 3(d), SE block consists of a global average pooling layer, two convolutional layers with the kernel size of 1*1, a ReLU and a Sigmoid activation functions. Given an input feature $F \in \mathbb{R}^{C,H,W}$, the design of SE block mainly consists of three steps:

1) Global average pooling (GAP) operator is used to squeeze global spatial information into a channel descriptor G, where $G \in \mathbb{R}^{C,1,1}$.

$$G = GAP(F) \in \mathbb{R}^{C,1,1} \tag{4}$$

2) The channel descriptor G is sequentially fed into a convolutional layer with the kernel size of C/r×1×1, a ReLU activation function, a convolutional layer with the kernel size of C×1×1 and a Sigmoid activation function to generate channel attention weights $W \in \mathbb{R}^{C,1,1}$ ranging from 0 to 1, which is the excitation operator for learning the dependence of each channel and adjusts the feature map according to the different dependence and where C is channel number and r is the compression ratio.

$$W = Sigmoid(Conv1 \times 1(\mathrm{ReLu}(\mathrm{Conv1} \times 1(G)))) \in \mathbb{R}^{C,1,1} \tag{5}$$

3) Finally, we multiply channel attention weights W by the original input feature F and do an element-wise summation with the original input feature F to obtain the final channel attention output $T \in \mathbb{R}^{C,H,W}$.

$$T = (F * W) + F \tag{6}$$

2.4. Loss functions

In our proposed network, the generative network takes the original unlabeled fundus images as input for images reconstruction, and the compound network outputs the results of classification

Biomedical Optics EXPRESS

and determines whether a fundus image is reconstructed by the generative network or not. Based on the analysis, the loss of our SSFC-ALN is divided into two parts: the loss of generative network and the loss of compound network.

The loss of generative network is formulated as:

$$L_G^{unsup} = \alpha * L_{ADV\_G} + (1-\alpha) * L_1(G) \tag{7}$$

$$L_{ADV\_G} = -\log \mathrm{D}(\mathrm{G}(\mathrm{x})) \tag{8}$$

$$L_1(G) = \frac{1}{W * H} \sum_{w=1}^{W} \sum_{h=1}^{H} |I_{w,h} - \mathrm{G}(I_{w,h})| \tag{9}$$

where L_{ADV_G} and $L_1(G)$ represent the adversarial loss and the image reconstruction loss of the generator in an unsupervised subset, respectively. α is a super-parameter referring to the weight of the unsupervised loss and is set to 0.001 in this study. W and H denote the size of an input fundus image, while $I_{w,h}$ indicates the image pixel value. Both W and H are 256 in this study.

The loss of compound network is formulated as follow:

$$L_{COM}^{semi-sup} = \beta * L_{CLS}^{sup} + \gamma * L_{ADV\_D\_U}^{unsup} + \delta * L_{ADV\_D\_R}^{unsup} \tag{10}$$

$$L_{CLS}^{sup} = -\frac{1}{m} \sum_{i=1}^{m} \sum_{k=1}^{K} I(t_i = k) \log(p(k|x_i)) \tag{11}$$

$$L_{ADV\_D\_U}^{unsup} = -\log D(x) \tag{12}$$

$$L_{ADV\_D\_R}^{unsup} = -\log(1 - D(G(x))) \tag{13}$$

where L_{CLS}^{sup} is the classification loss of classifier (K categories) in supervised subset, $L_{ADV\_D\_U}^{unsup}$ and $L_{ADV\_D\_R}^{unsup}$ are the adversarial loss of discriminator in unsupervised subset. β, γ and δ are super-parameters and all of them are set to $\frac{1}{3}$ in our experiments. m is the number of samples in per mini-batch, t_i denotes the class label of image x_i. I$(\cdot)$ is an indicator function, which equals one if t_i is equal to k.

3. Experiments and results

In this section, we first introduce the experimental dataset in detail. Then, the experimental setup will be described, including the parameter settings in the training phase and evaluation metrics in the testing phase. Finally, we will give the detailed experimental results and the corresponding analysis.

3.1. Dataset

In this study, Guangzhou Women and Children's Medical Center provided fundus images used for ROP zoning. These images with a resolution of $640 \times 480 \times 3$ were taken using RetCam3 camera by professional technicians between 2012 and 2015. The collection and analysis of image data were approved by the Institutional Review Board of Guangzhou Women and Children's Medical Center and adhered to the tenets of the Declaration of Helsinki. An informed consent was obtained from the guardians of each subject to perform all the imaging procedures. The gestation age varies from 26 to 41 weeks, with a mean value of 32 weeks. Fifty percent of infants' gestation age is under 32 weeks and 42% of the infants' birth weight is less than 1500 grams.

A total of 1013 fundus images of 108 patients from 192 examinations were labeled by a team of two attending ophthalmologists with over three years of ROP clinical experience and one experienced ROP specialist with more than fifteen years of ROP clinical experience from

Biomedical Optics EXPRESS

Guangzhou Women and Children Medical Center. The labeling of ROP zoning is based on the symptoms described in Table 1, and only the data with consistent results were used to evaluate the proposed network. It is statistically found that the category distribution is unbalanced, where most ROP data are in zone II and III and the data of zone I is relatively few. The possible reason is that effective treatment will be carried out before the disease progresses to the severest zone I in most cases. To evaluate the effectiveness of the proposed method, a 3-fold cross validation strategy is adopted. The training set and testing set of 3-fold are shown in the Table 2, which are randomly divided according to the examination of left and right eyes of each patient. In addition, we also collected 1317 fundus images of 105 premature infants as the unlabeled data set of this study. To reduce the computational cost and eliminate the effects of different scales and illuminations, all fundus images are downsampled to $256 \times 256 \times 3$ using bilinear interpolation and normalized to [1].

Table 2. Dataset used for training and testing the proposed method in this study.

Fold	Training			Testing		
	Zone I	Zone II	Zone III	Zone I	Zone II	Zone III
1	103	351	254	46	150	109
2	100	351	254	49	160	109
3	104	342	254	45	159	109

3.2. Experimental setup

3.2.1. Parameter setting

The proposed SSFC-ALN framework is implemented based on the PyTorch platform. We use a NVIDIA Tesla K40 GPU with 12GB memory to train the model with back-propagation algorithm by minimizing the loss function as shown in Eqs. (7) and (10). Adam is used as the optimizer to minimize the loss functions. Both initial learning rate and weight decay are set to 0.0001 to optimize the network. The batch size and epoch are set to 16 and 400, respectively. The compression ratio r is set to 16 in our study. In addition, the compound network updates twice as frequently as the generator in our proposed method. During training, all networks are trained with identical optimization schemes and we save the best model on validation set.

3.2.2. Evaluation metrics

Considering the category imbalance of the dataset shown in Table 2 and to fully and fairly evaluate the classification performance of different methods, four common classification metrics including weighted recall (W_R), weighted precision (W_P), weighted F1 score (W_F1) and Kappa index [53,54] are introduced to evaluate the ROP zoning performance.

3.3. Comparison experiments

We validate the proposed method on 1013 fundus images of 108 patients with a 3-fold cross validation strategy. Table 3 shows the quantitative results of different methods. As can be observed from Table 3, we compare our proposed method with other excellent supervised and semi-supervised CNN based classification networks, including ResNet34 [43], ResNet50 [43], ResNext50 [55], ResNext101 [55], InceptionV4 [44], DenseNet169 [41], SE_Resnet50 [35], SE_ResNext50 [35], EfficientNetB2 [56], ImprovedGAN [57] and Pix2PixGAN [32]. For the sake of fairness, DenseNet121 pre-trained on ImageNet is used as the feature extractor of the discriminator and classifier of ImprovedGAN and Pix2PixGAN. For the generator of the ImprovedGAN, we use 200-dimension vectors as input and add several transposed convolutional

Biomedical Optics EXPRESS

layers and batch normalized layers to the original version in order to generate $256 \times 256 \times 3$ fundus images [32]. The generator of Pix2PixGAN is U-shape encoder-decode architecture, which is same as the original version in [29]. Especially, for generators and discriminators, the training strategies of Pix2PixGAN and ImprovedGAN are the same as their original versions, and the training strategies of their classifiers are the same as the proposed network. In addition, to verify the effectiveness of the proposed FCM and semi-supervised adversarial learning and training strategies and to explore the influence of labeled data size, a series of ablation experiments are conducted. For convenience, we call the basic DenseNet121 pre-trained on ImageNet as the Baseline method.

Table 3. ROP zoning results of different methods.

Methods	W_R	W_P	W_F1	Kappa	Parameters (M)
ResNet34 [35]	0.7895 ± 0.0038	0.7986 ± 0.0066	0.7874 ± 0.0092	0.7434 ± 0.0417	21.2862
ResNet50 [35]	0.8453 ± 0.0150	0.8504 ± 0.0157	0.8451 ± 0.0143	0.8141 ± 0.0140	23.5142
ResNext50 [47]	0.8000 ± 0.0255	0.8054 ± 0.0247	0.7982 ± 0.0290	0.7523 ± 0.0556	22.9861
ResNext101 [47]	0.8219 ± 0.0414	0.8330 ± 0.0315	0.8179 ± 0.0495	0.7659 ± 0.0915	42.1349
InceptionV4 [36]	0.7061 ± 0.0155	0.7078 ± 0.0180	0.7004 ± 0.0121	0.6254 ± 0.0404	41.1474
DenseNet169 [32]	0.8509 ± 0.0467	0.8535 ± 0.0447	0.8486 ± 0.0502	0.8263 ± 0.0677	12.4895
SE_ResNet50 [31]	0.8518 ± 0.0640	0.8617 ± 0.0587	0.8468 ± 0.0709	0.8005 ± 0.1298	26.0452
SE_ResNext50 [31]	0.8222 ± 0.0489	0.8304 ± 0.0428	0.8210 ± 0.0493	0.7892 ± 0.0594	25.5170
EfficientNetB2 [48]	0.7813 ± 0.0249	0.7940 ± 0.0172	0.7783 ± 0.0297	0.7164 ± 0.0580	7.7095
ImprovedGAN [49]	0.8771 ± 0.0326	0.8832 ± 0.0363	0.8754 ± 0.0407	0.8333 ± 0.0594	22.4309
Pix2PixGAN [29]	0.8666 ± 0.0225	0.8710 ± 0.0226	0.8657 ± 0.0228	0.8283 ± 0.0521	64.1248
ResNet50 + ALN	0.8988 ± 0.0396	0.9006 ± 0.0378	0.8985 ± 0.0402	0.8711 ± 0.0767	37.1180
DenseNet169 + ALN	0.8884 ± 0.0534	0.8909 ± 0.0508	0.8985 ± 0.0550	0.8097 ± 0.1072	26.5249
Baseline	0.7800 ± 0.0198	0.7911 ± 0.0123	0.7769 ± 0.0252	0.7226 ± 0.0574	**6.9569**
SSFC-ALN	**0.9103 ± 0.0275**	**0.9155 ± 0.0278**	**0.9097 ± 0.0266**	**0.8919 ± 0.0446**	20.9923

Firstly, compared to Baseline, the performance of the proposed SSFC-ALN has been greatly improved, which improves the W_R, W_P, W_F1 and Kappa by 16.71%, 15.72%, 17.09% and 23.43%, respectively. Then, compared with other state-of-the-art supervised classification networks, the proposed method gets an overall improvement in terms of all metrics with comparable or less model complexity. For example, compared to the second best supervised learning classification network (SE_ResNet50), the W_R, W_P, W_F1 and Kappa of the proposed method increase from 0.8518, 0.8617, 0.8468 and 0.8005 to 0.9103, 0.9155, 0.9097 and 0.8919, respectively. In addition, compared to ResNet34, which has the comparable model complexity, our method has also made great improvement. The results show the effectiveness of semi-supervised learning. Similarly, ImprovedGAN and Pix2PixGAN also achieve great improvement by introducing semi-supervised learning. It is worth noting that the proposed SSFC-ALN has better performance than the above two semi-supervised GAN based methods and the model complexity is less, which may benefit from the appropriate optimization strategies adopted in the adversarial learning of generator and discriminator in this study. Notably, it can be seen from Table 3 that the zoning performance of "Baseline + FCM" was lower than ResNet50, DenseNet169 and SE_ResNet50. To further prove the advantage of selected compound network (Baseline + FCM), we also conduct the comparison experiments, of which we use ResNet50 and DenseNet169 as compound network (ResNet50 + ALN and DenseNet169 + ALN), respectively. There are two findings from Table 3. First, ResNet50 and DenseNet169 with semi-supervised adversarial learning outperform the supervised ResNet50 and DenseNet169, which further prove

Biomedical Optics EXPRESS

Table 4. Ablation study of FCM and semi-supervised adversarial learning.

Methods	W_R	W_P	W_F1	Kappa	Parameters (M)
Baseline	0.7800 ± 0.0198	0.7911 ± 0.0123	0.7769 ± 0.0252	0.7226 ± 0.0574	**6.9569**
Baseline + FCM	**0.8238 ± 0.0097**	**0.8377 ± 0.0239**	**0.8212 ± 0.0060**	**0.7870 ± 0.0083**	7.3885
Baseline + DANet	0.7973 ± 0.0348	0.8050 ± 0.0366	0.7956 ± 0.0369	0.7428 ± 0.0781	7.3451
Baseline + SE	0.8088 ± 0.0171	0.8175 ± 0.0130	0.8072 ± 0.0192	0.7607 ± 0.0427	6.9989
Baseline + DANet1	0.8184 ± 0.0167	0.8254 ± 0.0183	0.8172 ± 0.0151	0.7597 ± 0.0168	7.3880
Baseline + ALN	**0.8945 ± 0.0361**	**0.9064 ± 0.0348**	**0.8984 ± 0.0395**	**0.8880 ± 0.0455**	20.5602

the effectiveness of semi-supervised adversarial learning. Second, the proposed SSFC-ALN has better performance than ResNet50 + ALN and DenseNet169 + ALN with less model parameters, which proves the effectiveness of Baseline + FCM used in this study.

3.4. Ablation experiments

3.4.1. Ablation experiment for FCM

To prove the effect of the proposed FCM, we have conducted the ablation experiments as shown in Table 4. As can be seen from Table 4, the proposed FCM embedded in the classification network (Baseline + FCM) with a small increase in the amount of model parameters has made improvement over the Baseline in terms of all four evaluation indicators. Compared with the Baseline, the W_R, W_P, W_F1 and Kappa increase from 0.7800, 0.7911, 0.7769 and 0.7226 to 0.8238, 0.8377, 0.8212 and 0.7870, respectively. In addition, we also compare the proposed FCM with DANet and SE block. As shown in Table 4, the proposed FCM outperforms DANet and SE block. In addition, to further explore the advantage of using the SE block as the channel attention module of FCM, we have conducted an experiment, of which SE block replaces the original channel attention module from DANet (named "Baseline + DANet1"). As can be observed from Table 4, Baseline + DANet1 outperforms the Baseline + DANet, which shows SE block as channel attention module is better than the original channel attention module from DANet in our ROP zoning task. The possible reason is that the original channel attention module from DANet directly calculate the channel attention map from the original features without any convolution and nonlinear operation, while SE module uses 1×1 convolution and ReLu activation function, which can improve the expression and fitting ability of the network. To further demonstrate the effectiveness of the proposed FCM, we apply the "class activation mapping" technology [58] to obtain the heat maps of fundus images with different ROP zones for the qualitative analysis, which calculates the last convolutional outputs and visualizes the focus of the network. We compare the visualization results of FCM-intergraded network (Baseline + FCM) with Baseline. Figure 4 illustrates the visualization results. As can be observed from Fig. 4, the proposed FCM can focus on the target object regions better than Baseline. Taking the images in the first column of Fig. 4 as an example, Baseline only focuses on the optic disc area but not the ROP-related pathology, while FCM-intergraded network (Baseline + FCM) focuses on these two areas. Benefiting from it, our method can explicitly exploit information from the learned area, which is discriminative for ROP zoning. In addition, the target class scores increase accordingly, which indicates the proposed FCM can recalibrate the intermediate feature and make good use of the information of the target area and aggregate features from it to further improve the classification performance in our task.

3.4.2. Ablation experiment for semi-supervised adversarial learning

In this study, we propose a semi-supervised classification framework based on adversarial learning. Previous studies [23–29] suggest that, compared with the common supervised classification

Biomedical Optics EXPRESS

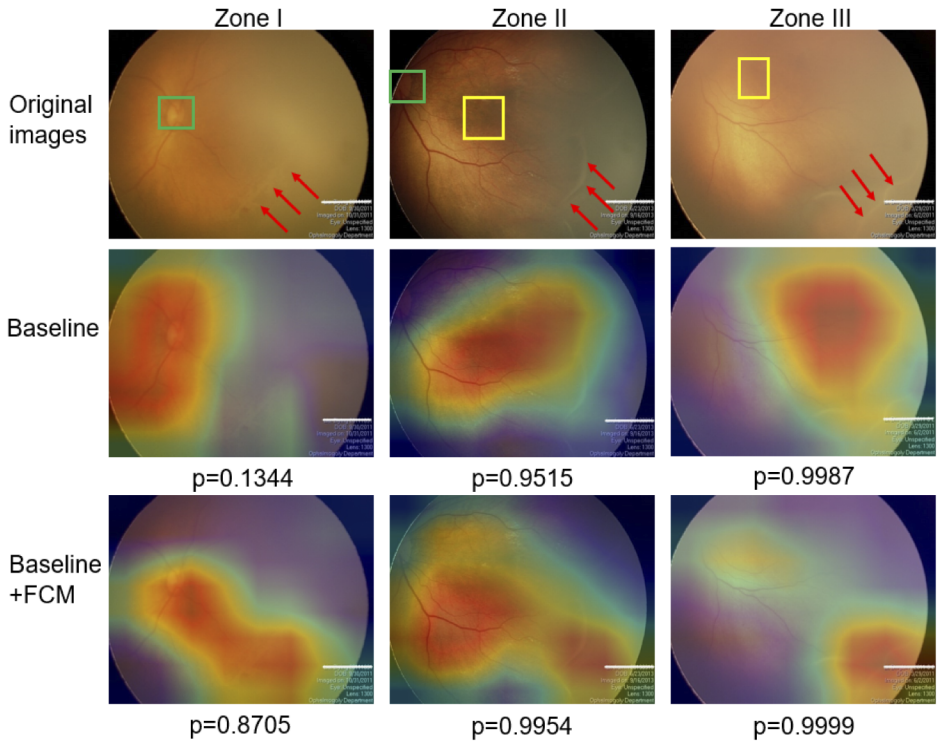

Fig. 4. CAM visualization results. The first row is the input images. The ground-truth label is shown on the top of each input image and p denotes the softmax score of each network for the ground-truth class. The optic disc and macula are in the green and yellow boxes respectively, and the ROP-related pathologies is indicated by the red marked arrows. The optic disc, macula, and ROP-related pathologies are the target areas that ophthalmologists are focusing on during ROP zoning recognition. The second row and the last row are the visualization results of Baseline and FCM-integrated network (Baseline + FCM).

network, the performance of the network can be improved by introducing semi-supervised learning based on GAN in the case of limited labeled data. To validate this viewpoint, we have also conducted ablation experiments with and without semi-supervised learning, which are shown in Table 4. As can be seen from Table 4, introducing semi- supervised learning based on GAN achieves a better classification performance in terms of all four metrics, with the W_R, W_P, W_F1 and Kappa increasing from 0.7800, 0.7911, 0.7769 and 0.7226 to 0.8945, 0.9064, 0.8984 and 0.8880, respectively. The results demonstrate the introduction of semi-supervised learning based on adversarial mechanism can reduce the dependence on labeled data and achieve better classification performance with limited labeled data in our ROP zoning task.

3.4.3. Ablation for training frequency

In this study, we introduce the idea of adversarial learning, trying to guide the training of classifier through the generative network. Based on the alternate training between the two models, the classifier model can better complete the classification task. At present, the mainstream scheme is to train two networks alternatively with 1:1 frequency. Considering that in different tasks different training frequency may bring different performances, we conduct several ablation experiments to explore the influence of training frequency of the generative network and compound network on the performance of ROP zoning. 'SSFC-ALN_2:1' denotes the generative network and the

Biomedical Optics EXPRESS

compound network are trained alternatively at a frequency of 2:1 and the meaning of the others is similar. The quantitative results are shown in Table 5. As can be seen from Table 5, with the decrease of the training frequency ratio of the generative network and the compound network, the performance of ROP zoning first improves and then decreases, and the best performance is obtained when the generator and the compound network are trained alternatively at a frequency of 1:2. The results indicate that appropriate training frequency can make the game between generative network and compound network more meaningful, so as to improve the overall performance of the classifier.

Table 5. Ablation study of training frequency on ROP zoning data in this paper

Methods	W_R	W_P	W_F1	Kappa
SSFC-ALN_2:1	0.8881 ± 0.0444	0.8923 ± 0.0451	0.8869 ± 0.0452	0.8297 ± 0.1299
SSFC-ALN _1:1	0.8934 ± 0.0381	0.8958 ± 0.0364	0.8922 ± 0.0390	0.8348 ± 0.1093
SSFC-ALN _1:2	**0.9103 ± 0.0275**	**0.9155 ± 0.0278**	**0.9097 ± 0.0266**	**0.8919 ± 0.0446**
SSFC-ALN _1:3	0.8976 ± 0.0496	0.9020 ± 0.0493	0.8962 ± 0.0506	0.8386 ± 0.1369

3.4.4. Ablation experiment for labeled data size

In this section, we explore the influence of labeled data size on the performance of ROP zoning. We randomly selected 20%, 40%, 60% and 80% fundus images from the whole labeled training set as the new labeled training set. The experimental results of different labeled data sizes are shown in Table 6, where 'Baseline_0.2' denotes Baseline method with 20% labeled training data and the meaning of the others is similar. There are three main findings from Table 6. First, with the increase of labeled data, the performance of supervised learning and semi-supervised learning has been improved. Second, under the same amount of labeled training data, the performance of semi-supervised learning based on adversarial learning has been greatly improved compared with the Baseline method. Finally, with the increase of labeled training fundus images, our method can achieve more improvements than Baseline method. The above results further indicate our method is effective and can reduce the dependence on labeled training data.

Table 6. Ablation study of labeled data size on ROP zoning data in this paper

Methods	W_R	W_P	W_F1	Kappa
Baseline_0.2	0.6967 ± 0.0445	0.7054 ± 0.0527	0.6957 ± 0.0486	0.6236 ± 0.0697
Baseline_0.4	0.7213 ± 0.0169	0.7451 ± 0.0193	0.7126 ± 0.0177	0.6104 ± 0.0304
Baseline _0.6	0.7564 ± 0.0278	0.7650 ± 0.0318	0.7558 ± 0.0273	0.6915 ± 0.0159
Baseline _0.8	0.7736 ± 0.0198	0.7782 ± 0.0218	0.7714 ± 0.0215	0.7032 ± 0.0451
Baseline _1.0	0.7800 ± 0.0198	0.7911 ± 0.0123	0.7769 ± 0.0252	0.7226 ± 0.0574
SSFC-ALN _0.2	0.7399 ± 0.1050	0.7482 ± 0.1280	0.7137 ± 0.1435	0.6082 ± 0.2307
SSFC-ALN_0.4	0.8472 ± 0.0340	0.8598 ± 0.0252	0.8452 ± 0.0360	0.8091 ± 0.0583
SSFC-ALN_0.6	0.8785 ± 0.0570	0.8942 ± 0.0426	0.8743 ± 0.0614	0.8500 ± 0.0870
SSFC-ALN_0.8	0.8946 ± 0.0333	0.8974 ± 0.0318	0.8937 ± 0.0348	0.8463 ± 0.0655
SSFC-ALN_1.0	**0.9103 ± 0.0275**	**0.9155 ± 0.0278**	**0.9097 ± 0.0266**	**0.8919 ± 0.0446**

Biomedical Optics EXPRESS

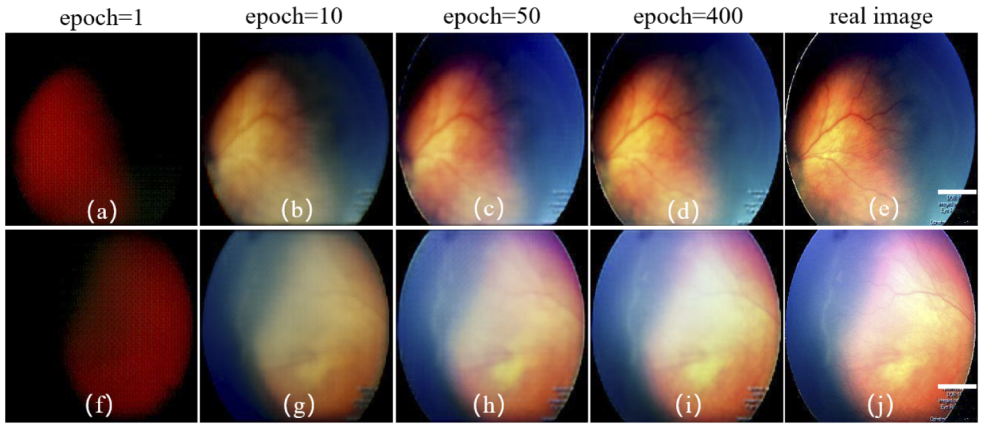

Fig. 5. The comparison of generated images and real images. (a)-(d) are the generated images corresponding to the real image (e) at different training epoch. Similarly, (f)-(i) are the generated images corresponding to the real image (j) at different training epoch.

4. Conclusion and discussions

The inadequacy of labeled data is a challenge for using deep learning based algorithms in medical image analysis. The main reasons are as follows: 1) the high-quality labeling process of medical images is very expensive because it depends on scarce medical expertise; 2) compared with natural problems, medical image acquisition is more difficult due to privacy problems. Actually, the collection of sufficient labeled ROP images is more difficult. In this paper, we propose a novel semi-supervised feature calibration adversarial learning network (SSFC-ALN) for 3-level ROP zoning. First, we propose a novel attention module named feature calibration module (FCM), which is embedded in the middle layer of the compound network, and can effectively calibrate the intermediate features from two dimensions of space and channel to improve the feature representation of the network. Then, to reduce the dependence on labeled data and make full use of unlabeled dataset, semi-supervised learning based on GAN is introduced, which can introduce the idea of adversarial learning mechanism to improve the performance of ROP zoning in the alternate adversarial training of the generative network and the compound network. To the best of our knowledge that it is the first time that semi-supervised learning is introduced into ROP zoning task, and good performance is achieved. Finally, appropriate optimization strategy is adopted and good classification performance is achieved.

The comprehensive experiments show the effectiveness of the proposed method in our ROP zoning task. Compared with other state-of-the-art supervised CNN-based methods, our proposed SSFC-ALN with similar or less complexity can adaptively focus on the related area of ROP zoning and significantly improve the accuracy of ROP zoning and the generalization ability of model. In addition, to further evaluate the performance of our proposed method, we also compare our SSFC-ALN with ImprovedGAN and Pix2PixGAN as shown in Table 3, which demonstrates that our proposed method has advantages in both accuracy and model complexity. Especially, the comparison between the generated images and real images is shown in Fig. 5. As can be observed from Fig. 5, with the increase of training epoch, the images generated by the generator are getting closer and closer to the original input images, but it is not completely consistent. It is worth noting that the purpose of this study is to obtain a good classifier (discriminator). As mentioned in Ref. [59], good semi-supervised classification performance and good generator cannot be obtained at the same time. It turns out that our practical observation is consistent with the theory of Ref. [59]. Therefore, in this study, it is not important whether the generated images

Biomedical Optics EXPRESS

are consistent with the real images. In fact, the important thing is that the use of generators for image reconstruction can introduce a large amount of unlabeled data, and at the same time establish an adversarial training mechanism with the compound network.

Although the proposed method has achieved good performance on existing ROP zoning datasets, there are still some limitations. Firstly, the evaluation of all the comparison algorithms and the proposed method are based on limited labeled data. More high-quality clinical labeled ROP zoning data should be collected to further validate the performance of the proposed method. In addition, biological methods and pathological analysis are not considered in this paper. Therefore, in the future, we will collect more high-quality labeled data, extend the proposed method to other ROP related analysis (such as the identification of AP-ROP, plus disease and five stages of ROP) and combine ROP artificial intelligence diagnosis method with biological methods and pathological analysis, aiming to comprehensively assist the ophthalmologist in clinical diagnosis and treatment of ROP.

Funding. National Natural Science Foundation of China (61622114, U20A20170); National Key Research and Development Program of China (2018YFA0701700).

Disclosures. The authors declare that there are no conflicts of interest related to this article.

Data availability. The dataset underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

References

1. J. Chen and L. E. H. Smith, "Retinopathy of prematurity," Angiogenesis **10**(2), 133–140 (2007).
2. S. J. Kim, A. D. Port, R. Swan, J. P. Campbell, R. V. P. Chan, and M. F. Chiang, "Retinopathy of prematurity: a review of risk factors and their clinical significance," Surv. Ophthalmol. **63**(5), 618–637 (2018).
3. D. K. VanderVeen, "Prevalence and course of strabismus in the first year of life for infants with prethreshold Retinopathy of Prematurity," Arch. Ophthalmol. **124**(6), 766–773 (2006).
4. J. Rao, D. Fan, D. Lin, H. Zhang, S. Ye, X. Luo, L. Wang, J. Yang, and M. Pang, "Trend and risk factors of low birth weight and macrosomia in south China, 2005–2017: a retrospective observational study," Scientific Reports. **8**(1), 3393–3400 (2018).
5. G. E. Quinn, G. Ying, E. F. Bell, P. K. Donohue, D. Morrison, L. A. Tomlinson, and G. Binenbaum, "Incidence and early course of retinopathy of prematurity: secondary analysis of the postnatal growth and retinopathy of prematurity (G-ROP) study," JAMA Ophthalmol. **136**(12), 1383–1389 (2018).
6. C. Wu, R. A. Petersen, and D. K. VanderVeen, "RetCam imaging for retinopathy of prematurity screening," J. Am. Assoc. Pediatric Ophthalmol. Strabismus **10**(2), 107–111 (2006).
7. Committee for the Classification of Retinopathy of Prematurity, "An international classification of retinopathy of prematurity," Arch. Ophthalmol.. **102**(8), 1130–1134 (1984).
8. T. Aaberg, "An international classification of retinopathy of prematurity: II. The classification of retinal detachment," Arch. Ophthalmol. **105**(7), 906–912 (1987).
9. International Committee for the Classification of Retinopathy of Prematurity, "The international classification of retinopathy of prematurity revisited," Arch. Ophthalmol.. **123**(7), 991–999 (2005).
10. D. K. Wallace, Z. Zhao, and S. F. Freedman, "A pilot study using "ROPtool" to quantify plus disease in retinopathy of prematurity," J. Am. Assoc. Pediatric Ophthalmol. Strabismus **11**(4), 381–387 (2007).
11. A. C. Esra, B. C. Veronica, C. J. Peter, B. Alican, K. C. Jayashree, P. Samir, J. Karyn, R. V. P. Chan, and O. Susan, "Computer-based image analysis for plus disease diagnosis in retinopathy of prematurity: performance of the "i-ROP" system and image features associated with expert diagnosis," Trans. Vis. Sci. Tech. **4**(6), 5–16 (2015).
12. D. E. Worrall, C. M. Wilson, and G. J. Brostow, "Automated retinopathy of prematurity case detection with convolutional neural networks," *International Workshop on Deep Learning in Medical Image Analysis* (2016), 68–76.
13. J. M. Brown, J. P. Campbell, A. Beers, K. Chang, S. Ostmo, R. V. P. Chan, J. Dy, D. Erdogmus, S. Loannidis, J. Kalpathy-Cramer, and M. F. Chiang, "Automated diagnosis of plus disease in retinopathy of prematurity using deep convolutional neural networks," JAMA Ophthalmol. **136**(7), 803–810 (2018).
14. J. Hu, Y. Chen, J. Zhong, R. Ju, and Z. Yi, "Automated analysis for retinopathy of prematurity by deep neural networks," IEEE Trans. Med. Imaging **38**(1), 269–279 (2019).
15. J. Wang, R. Ju, Y. Chen, L. Zhang, and J. Hu, "Automated retinopathy of prematurity screening using deep neural networks," EBioMedicine **35**(1), 361–368 (2018).
16. Y. Zhang, L. Wang, Z. Wu, J. Zeng, Y. Chen, R. Tain, J. Zhao, and G. Zhang, "Development of an automated screening system for retinopathy of prematurity using a deep neural network for wide-angle retinal images," IEEE Access **7**(1), 10232–10241 (2018).
17. R. Zhang, J. Zhao, G. Chen, T. Wang, G. Zhang, and B. Lei, "Aggressive posterior retinopathy of prematurity automated diagnosis via a deep convolutional network," In *International Workshop on Ophthalmic Medical Image Analysis* (Springer, 2019), 165–172.

Biomedical Optics EXPRESS

18. G. Chen, J. Zhao, R. Zhang, T. Wang, G. Zhang, and B. Lei, "Automated stage analysis of retinopathy of prematurity using joint segmentation and multi-instance learning," In *International Workshop on Ophthalmic Medical Image Analysis* (Springer, 2019), 173–181.
19. Y. Peng, W. Zhu, F. Chen, D. Xiang, and X. Chen, "Automated retinopathy of prematurity screening using deep neural network with attention mechanism," In *Medical Imaging 2020: Image Processing* (2020), 1131321–1131327.
20. Y. Peng, W. Zhu, F. Chen, and X. Chen, "Automated zone recognition for retinopathy of prematurity using deep neural network with attention mechanism and deep supervision strategy," In *Medical Imaging 2021: Image Processing*, International Society for Optics and Photonics (2021), 115961–115968.
21. Y. Peng, W. Zhu, Z. Chen, M. Wang, L. Geng, K. Yu, Y. Zhou, T. Wang, D. Xiang, F. Chen, and X. Chen, "Automatic staging for retinopathy of prematurity with deep feature fusion and ordinal classification strategy," IEEE Trans. Med. Imaging **40**(7), 1750–1762 (2021).
22. J. Zhao, B. Lei, Z. Wu, Y. Zhang, and G. Zhang, "A deep learning framework for identifying zone I in RetCam images," IEEE Access. **7**(1), 103530 (2019).
23. R. Agrawal, S. Kulkarni, R. Walambe, and K. Kotecha, "Assistive framework for automatic detection of all the zones in retinopathy of prematurity using deep learning," Journal of Digital Imaging. **34**(2), 1–16 (2021).
24. E. Alvarez, M. Wakakura, Z. Khan, and G. N. Dutton, "The disc-macula distance to disc diameter ratio: a new test for confirming optic nerve hypoplasia in young children," J. Pediatr Ophthalmol. Strabismus **25**(3), 151–154 (1988).
25. D. D. Silva, K. D. Cocker, G. Lau, S. T. Clay, A. R. Fielder, and M. J. Moseley, "Optic disk size and optic disk-to-fovea distance in preterm and full-term infants," Invest. Ophthalmol. Vis. Sci. **47**(11), 4683–4686 (2006).
26. X. Yi, E. Walia, and P. Babyn, "Unsupervised and semi-supervised learning with categorical generative adversarial networks assisted by Wasserstein distance for dermoscopy image classification," arXiv preprint arXiv:1804.03700 (2018).
27. S. Wang, X. Wang, Y. Shen, Z. Yang, M. Gan, and B. Lei, "Diabetic retinopathy diagnosis using multichannel generative adversarial network with semi-supervision," IEEE Trans. Automat. Sci. Eng. **18**(2), 574–585 (2021).
28. B. Lecouat, K. Chang, C. Foo, B. Unnikrishnan, J. M. Brown, H. Zenati, A. Beers, V. Chandrasekhar, J. Kalpathy-Cramer, and P. Krishnaswamy, "Semi-supervised deep learning for abnormality classification in retinal images," arXiv preprint arXiv:1812.07832 (2018).
29. S. Liu, J. Xin, J. Wu, and P. Shi, "Semi-supervised adversarial learning for diabetic retinopathy screening," *Ophthalmic Medical Image Analysis: 6th International Workshop*, OMIA (2019), 60–68.
30. G. Yang, S. Yu, H. Dong, G. Slabaugh, P. L. Dragotti, X. Ye, F. Liu, S. Arridge, J. Keegan, Y. Guo, and D. Firmin, "DAGAN: Deep de-aliasing generative adversarial networks for fast compressed sensing MRI reconstruction," IEEE Trans. Med. Imaging **37**(6), 1310–1321 (2018).
31. Z. Li, Y. Wang, and J. Yu, "Reconstruction of thin-slice medical images using generative adversarial network," *International Workshop on MachineLearning in Medical Imaging* (Springer, 2017), 325–333.
32. P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2017), 5967–5976.
33. V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," *Advances in Neural Information Processing Systems* (2014), 1–12.
34. J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2019), 3146–3154.
35. J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2018), 7132–7141.
36. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," Advances in neural information processing systems. **3**(11), 1–9 (2014).
37. M. Mirza and S. Osindero, "Conditional generative adversarial nets," arXiv preprint arXiv:1411.1784, 2672-2680 (2014).
38. Y. Ma, X. Chen, W. Zhu, X. Cheng, D. Xiang, and F. Shi, "Speckle noise reduction in optical coherence tomography images based on edge-sensitive cGAN," Biomed. Opt. Express **9**(11), 5129–5146 (2018).
39. H. Jiang, X. Jian, F. Shi, Y. Ma, D. Xiang, L. Ye, J. Su, Z. Li, Q. Chen, Y. Hua, X. Xu, W. Zhu, and Y. Fan, "Improved cGAN based linear lesion segmentation in high myopia ICGA images," Biomed. Opt. Express **10**(5), 2355–2366 (2019).
40. M. Wang, W. Zhu, K. Yu, Z. Chen, F. Shi, Y. Zhou, Y. Ma, Y. Peng, D. Bao, S. Feng, L. Ye, D. Xiang, and X. Chen, "Semi-supervised capsule cGAN for speckle noise reduction in retinal OCT images," IEEE Trans. Med. Imaging **40**(4), 1168–1183 (2021).
41. G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (IEEE, 2017), 4700–4708.
42. S. Pan and Q. Yang, "A survey on transfer learning," IEEE Transactions on Knowledge and Data Engineering. 22(10), 1345–1359 (2010).
43. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2016), 770–778.
44. C. Szegedy, S. Loffe, V. Vincent, and A. Alex, "Inception-v4, inception-resnet and the impact of residual connections on learning," arXiv preprint arXiv:1602.07261 (2016).

Biomedical Optics EXPRESS

45. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scaleimage recognition," arXiv preprint arXiv:1409.1556 (2014).
46. X. Li, W. Wang W, X Hu, and J. Yang, "Selective kernel networks," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2020), 1–12.
47. X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2018), 7794–7803.
48. F. Wang, M. Jiang, C. Qian, S. Yang, and X. Tang, "Residual attention network for image classification," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2017) 1–10.
49. S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," In *Proceedings of the European Conference on Computer Vision* (2018), 3–19.
50. T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, "The application of two-level attention models in deep convolutional neural network for fine-grained image classification," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2015), 2–10.
51. B. Zhao, X. Wu, J. Feng, Q. Peng, S. Yan, and B. Zhao, "Diversified visual attention networks for fine-grained object classification," IEEE Trans. Multimedia **19**(6), 1245–1256 (2017).
52. P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," arXiv preprint arXiv:1803.02155 (2018).
53. J. Carletta, "Assessing agreement on classification tasks: the kappa statistic," arXiv preprint cmp-lg/9602004 (1996).
54. M. L. McHugh, "Interrater reliability: the kappa statistic," Biochem. Med. **22**(3), 276–282 (2012).
55. S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2017), 1492–1500.
56. M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," arXiv, preprint arXiv:1905.11946 (2019).
57. T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANS," *Conference and Workshop on Neural Information Processing Systems* (2016), 2234–2242.
58. B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2016), 2921–2929.
59. Z. Dai, Z. Yang, F. Yang, W. William, and R. Salakhutdinov, "Good semi-supervised learning that requires a bad GAN," arXiv preprint arXiv:1705.09783 (2017).